Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means ☆

Frank Nielsen *

Sony Computer Science Laboratories, Inc., 3-14-13 Higashi Gotanda, Shinagawa-Ku, Tokyo 141-0022, Japan
Joseph-Louis Lagrange laboratory (CNRS, OCA), Univ. Nice Sophia-Antipolis, UFR Sciences, Parc Valrose, Building H. Fizeau, 06108 NICE Cedex 2, France

ABSTRACT

Bayesian classification labels observations based on given prior information, namely class-*a priori* and class-conditional probabilities. Bayes' risk is the minimum expected classification cost that is achieved by the Bayes' test, the optimal decision rule. When no cost incurs for correct classification and unit cost is charged for misclassification, Bayes' test reduces to the maximum *a posteriori* decision rule, and Bayes risk simplifies to Bayes' error, the probability of error. Since calculating this probability of error is often intractable, several techniques have been devised to bound it with closed-form formula, introducing thereby measures of similarity and divergence between distributions like the Bhattacharyya coefficient and its associated Bhattacharyya distance. The Bhattacharyya upper bound can further be tightened using the Chernoff information that relies on the notion of best error exponent. In this paper, we first express Bayes' risk using the total variation distance on scaled distributions. We then elucidate and extend the Bhattacharyya and the Chernoff upper bound mechanisms using generalized weighted means. We provide as a byproduct novel notions of statistical divergences and affinity coefficients. We illustrate our technique by deriving new upper bounds for the univariate Cauchy and the multivariate *t*-distributions, and show experimentally that those bounds are not too distant to the computationally intractable Bayes' error.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction: hypothesis testing, divergences and affinities

1.1. Hypothesis testing

Consider the following fundamental *binary hypothesis testing* problem[1]: Let $X_1, \ldots, X_n$ be n identically and independently distributed (IID) random variables following distribution Q with support $\mathcal{X}$. We consider two (simple) hypotheses:

$$H_1 : Q \sim P_1 (\text{null hypothesis}), \tag{1}$$

$$H_2 : Q \sim P_2 (\text{alternative hypothesis}) \tag{2}$$

and we design a *test* $g(X_1, \ldots, X_n) : \mathcal{X}^n \to \{1, 2\}$ to decide which hypothesis to select. The decision region $R_1 \subseteq \mathcal{X}^n$ corresponds to

the set of sequences $X^n = (X_1, \ldots, X_n)$ mapped to H_1, and the decision region $R_2 = R_1^c$ is the complementary region.

To illustrate this setting, consider for example the task of distinguishing a texture [21], modeled by edgelets[2] X^n centered at 2D image lattice positions, from two textures T_1 and T_2, given by their respective edgelet probability distributions P_1 and P_2 (assuming the IID hypothesis). In practice, we *observe* a *texture sample*, that is a data set $x^n = (x_1, \ldots, x_n)$ sampled from X^n.

There are two kinds of error [5] associated with any test:

- Type I error (misclassification when the true hypothesis is H_1): $\epsilon_1(n) = \Pr(g(X_1, \ldots, X_n) = 2 | H_1)$, and
- Type II error (misclassification when the true hypothesis is H_2): $\epsilon_2(n) = \Pr(g(X_1, \ldots, X_n) = 1 | H_2)$.

In target/noise detection theory [5], a test is called a *detector*, and those type I and type II errors are respectively called probability of *false alarm* and probability of *miss*.

There are two main approaches for hypothesis testing that have been developed in the literature [5,7]: The first Neyman–Pearson

---

[1] We refer the reader to the textbooks [5,6] for an information-theoretic background based on the method of types, and to the textbook [7] for the Bayesian setting often met in pattern recognition.

[2] An edgelet is a small line segment with slope quantized to take d possible directions.

approach seeks to minimize the probability of miss given the probability of false alarm, without any prior information for the hypothesis. The second Bayesian approach makes use of prior information on class-*a priori* and class-conditional probabilities for the hypothesis. We concisely review the links between hypothesis testing and statistical distances between distributions for the first non-Bayesian approach in Section 1.2, and the links between hypothesis testing and statistical similarities between distributions for the second Bayesian approach in Section 1.3.

1.2. Statistical divergences in hypothesis testing

The first approach asks to minimize the probability of miss ϵ_2 (*false negative*) given the probability of false alarm ϵ_1 (*false positive*):

$$\min_{\epsilon_2} \epsilon_1 \leqslant \varepsilon, \tag{3}$$

where ε is a prescribed error threshold. In the literature, the *significance level* (or size) of a test is ϵ_1 and the *power* of a test is $1 - \epsilon_2$. Thus we seek to maximize the *power of a test* given a prescribed significance level. A key result is the Neyman–Pearson lemma [5] which states the optimality of the *Likelihood Ratio Test* (LRT) (or equivalently its log-likelihood ratio):

$$\Lambda(X_1, \ldots, X_n) = \log \frac{P_1(X_1, \ldots, X_n)}{P_2(X_1, \ldots, X_n)} = \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_2(X_i)} \leqslant \lambda, \tag{4}$$

to reject H_1 in favor of H_2 (with $\Pr(\Lambda(X_1, \ldots, X_n) \leqslant \lambda | H_1) = \epsilon_1$ and $\lambda = \lambda(\varepsilon)$). Note that the larger the log-likelihood ratio, the more probable the sequence X^n comes from P_1 (and the more likely hypothesis H_1). From the IDD assumption, we better analyze sequences via the *method of types* [21,6]: The type $h(x^n)$ is the empirical probability distribution of elements of $\mathcal{X}$ (say, a discrete alphabet with d letters) met in x^n. That is, for alphabet $\mathcal{X} = \{E_1, \ldots, E_d\}$, the type $h(x^n)$ of a sample sequence is the frequency empirical histogram of elements: $h(x^n) = (h_1(x^n), \ldots, h_d(x^n))$, where $h_i(x^n) = \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j, E_i} = \frac{\#\{x_j = E_i \mid j \in \{1, \ldots, n\}\}}{n}$. Observe that although there are d^n distinct sequences of length n (that is, exponential in n), there is only a polynomial number of types (bounded by $(n+1)^d$ since each of the d elements E_i of $\mathcal{X}$ has counting number $n_i = nh_i(X^n)$, an integer between 0 and n).

The log-likelihood ratio of Eq. (4) rewrites as:

$$\Lambda(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_2(x_i)} = \sum_{j=1}^{d} n h_j(x^n) \log \frac{P_1(E_j)}{P_2(E_j)}. \tag{5}$$

Observe that the log-likelihood ratio can be conveniently written as an inner product between two histograms $h(x^n)$ and $A_{12} = \left(\log \frac{P_1(E_1)}{P_2(E_1)}, \ldots, \log \frac{P_1(E_d)}{P_2(E_d)} \right)$: $\Lambda(x_1, \ldots, x_n) = n \langle h(x^n) | A_{12} \rangle$. In [21], the inner product $\langle h(x^n) | A_{12} \rangle$ is called the *reward* of sequence sample x^n.

The *expected average log-likelihood ratios* with respect to P_1 and P_2 are:

$$\frac{1}{n} E_{P_1}[\Lambda(X_1, \ldots, X_n)] = \sum_{j=1}^{d} P_1(E_j) \log \frac{P_1(E_j)}{P_2(E_j)} = \mathrm{KL}(P_1 : P_2), \tag{6}$$

$$\frac{1}{n} E_{P_2}[\Lambda(X_1, \ldots, X_n)] = \sum_{j=1}^{d} P_2(E_j) \log \frac{P_1(E_j)}{P_2(E_j)} = -\mathrm{KL}(P_2 : P_1), \tag{7}$$

where $\mathrm{KL}(P_1 : P_2) = \sum_{i=1}^{d} P_1(E_i) \log \frac{P_1(E_i)}{P_2(E_i)}$ denotes the *Kullback–Leibler divergence* between distributions P_1 and P_2. The difference between the two expected average log-likelihood ratio is the *Jeffreys divergence* $J(P_1, P_2) = \mathrm{KL}(P_1 : P_2) + \mathrm{KL}(P_2 : P_1)$, that symmetrizes the Kullback–Leibler divergence. It follows that those Kullback–Leibler and Jeffreys measures can be interpreted as *measures of separability*,

that is distances between distributions. Note that the KL distance is asymmetric: $\mathrm{KL}(P_1 : P_2) \neq \mathrm{KL}(P_2 : P_1)$. The KL and J distances are not metric because they violate the triangular inequality [5].

The probability that a sequence sample x_1^n from P_1 has lower reward than a sequence sample x_2^n from P_2 is bounded by (see [21, Theorem 2]):

$$(n+1)^{-d^2} 2^{-nB(P_1, P_2)} \leqslant \Pr(\langle h(x_1^n) | A_{12} \rangle \leqslant \langle h(x_2^n) | A_{12} \rangle)$$
$$\leqslant (n+1)^{d^2} 2^{-nB(P_1, P_2)}, \tag{8}$$

where $B(P_1, P_2)$ denotes the *Bhattacharrya divergence*:

$$B(P_1, P_2) = -\log \sum_{j=1}^{d} \sqrt{P_1(E_j)} \sqrt{P_2(E_j)}. \tag{9}$$

Although the Neyman–Pearson lemma [5] characterizes the optimal decision test, it does not specify the threshold λ. The false alarm error ϵ_1 and miss error ϵ_2 probabilities decay exponentially as the sample size n increases (see [21, Theorem 1]). Thus in the asymptotic regime, we are rather interested in characterizing the *error exponents* defined as the rate of that exponential decay:

$$\alpha = \lim_{n \to \infty} -\frac{1}{n} \log_2 \epsilon_1(n), \quad \beta = \lim_{n \to \infty} -\frac{1}{n} \log_2 \epsilon_2(n), \tag{10}$$

where $\epsilon_1(n) = \Pr(g(X^n) = 2 | H_1)$ and $\epsilon_2(n) = \Pr(g(X^n) = 1 | H_2)$ (and $\epsilon_1(n) \approx 2^{-n\alpha}$ and $\epsilon_2(n) \approx 2^{-n\beta}$). It turns out that when minimizing the asymptotic rate of *misclassification error* $P_e = \epsilon_1 + \epsilon_2$, called the probability of error, the optimal threshold is $\lambda = 0$, and the error rate is the Chernoff information [5,15]:

$$C(P_1, P_2) = \min_{\alpha \in [0,1]} B_\alpha(P_1 : P_2), \tag{11}$$

where B_α denotes the *skewed Bhattacharrya divergence*:

$$B_\alpha(P_1 : P_2) = -\log \sum_{j=1}^{d} P_1(E_j)^\alpha P_2(E_j)^{1-\alpha}, \tag{12}$$

generalizing the Bhattacharrya divergence: $B(P_1, P_2) = B_{\frac{1}{2}}(P_1 : P_2)$.

Those notions of statistical divergences can be extended to continuous distributions by replacing the discrete sum by an integral (and interpreted $\mathcal{X}$ as a continous alphabet).

We now consider the Bayesian paradigm in hypothesis testing, and show how to bound the probability of misclassification error using statistical similarity measures.

1.3. Statistical similarities in hypothesis testing

The Bayesian framework of hypothesis testing assumes that we are given prior beliefs over the probabilities of the two hypothesis, and we seek to minimize the *expected probability of error* (also called *error probability*): $P_e = \epsilon_1 \Pr(H_1) + \epsilon_2 \Pr(H_2)$. In this setting, both the *class a priori* ($w_i > 0$) and the *class conditional* probabilities (p_i) are known beforehand (or estimated from a training labeled sample [7]). Let q_1 and q_2 be the *a posteriori* probabilities derived from Bayes theorem:

$$q_i(x) = \frac{w_i p_i(x)}{p(x)}, \tag{13}$$

where $p(x)$ is the mixture density $p(x) = w_1 p_1(x) + w_2 p_2(x)$ (and $w_1 + w_2 = 1$). Let $C = [c_{ij}]$ be the 2×2 *design matrix*, with c_{ij} denoting the *cost* of deciding $x \in C_i$ when $x \in C_j$, with $1 \leqslant i, j \leqslant 2$. Furthermore, denote by

$$r_1(x) = c_{11} q_1(x) + c_{12} q_2(x) \tag{14}$$

and

$$r_2(x) = c_{21} q_1(x) + c_{22} q_2(x), \tag{15}$$

the respective *conditional costs* of deciding $x \in C_i$, for $i \in \{1, 2\}$. To classify x, consider the decision rule:

$$r_1(x) \underset{C_2}{\overset{C_1}{\lessgtr}} r_2(x). \tag{16}$$

The conditional cost of this decision rule is:

$$r(x) = \min(r_1(x), r_2(x)). \tag{17}$$

Bayes error B_e (see [7], p. 57) is defined as the *expected* cost of this decision rule:

$$B_e = E_p[r(x)], \tag{18}$$

$$= \int p(x) \min(r_1(x), r_2(x)) \mathrm{d}x, \tag{19}$$

$$= \int_{R_1} (c_{11} w_1 p_1(x) + c_{12} w_2 p_2(x)) \mathrm{d}x + \int_{R_2} (c_{21} w_1 p_1(x) + c_{22} w_2 p_2(x)) \mathrm{d}x, \tag{20}$$

where $R_1 = \{x \mid r_1(x) \leqslant r_2(x)\}$ and $R_2 = \{x \mid r_2(x) \leqslant r_1(x)\}$ are the *decision regions* induced by the decision rule of Eq. (16). Thus Bayes test for minimum cost writes as:

$$(c_{12} - c_{22}) w_2 p_2(x) \underset{C_2}{\overset{C_1}{\gtrless}} (c_{21} - c_{11}) w_1 p_1(x), \tag{21}$$

or equivalently:

$$\frac{p_1(x)}{p_2(x)} \underset{C_2}{\overset{C_1}{\gtrless}} \frac{w_2(c_{12} - c_{22})}{w_1(c_{21} - c_{11})}. \tag{22}$$

The term $l(x) = \frac{p_1(x)}{p_2(x)}$ is called the *likelihood ratio*. It is equivalent and often mathematically simpler to consider the test using the log-likelihood ratio (e.g., think of the multivariate Gaussian class-conditional probabilities):

$$\log p_1(x) - \log p_2(x) \underset{C_2}{\overset{C_1}{\lessgtr}} \log \frac{w_2(c_{12} - c_{22})}{w_1(c_{21} - c_{11})} \tag{23}$$

The function $h(x) = \log p_1(x) - \log p_2(x) - \frac{w_2(c_{12} - c_{22})}{w_1(c_{21} - c_{11})}$ is called the *discriminant function*.

For *symmetrical cost* function $c_{21} - c_{11} = c_{12} - c_{22}$, the expected cost is called the *probability of error* P_e. For the probability of error, we do not incur a cost for correctly classifying and pay a unit cost for misclassification. The design matrix for the probability of error is therefore:

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{24}$$

As stated earlier, the probability of error can be decomposed as the sum of two misclassification costs:

$$P_e = w_1 \epsilon_1 + w_2 \epsilon_2, \tag{25}$$

with

$$\epsilon_1 = \int_{R_2} p_1(x) \mathrm{d}x, \quad \epsilon_2 = \int_{R_1} p_2(x) \mathrm{d}x, \tag{26}$$

with $R_1 = \{w_2 p_2(x) \leqslant w_1 p_1(x)\}$ and $R_2 = \{x \mid w_1 p_1(x) \leqslant w_2 p_2(x)\}$.

In practice, Bayes error and the probability of error are quite tricky to calculate as we need to compute integrals on decision regions R_1 and R_2. Even if those domains can be expressed simply, say for Gaussians, it is often intractable to compute analytically those integrals (e.g., for multivariate class-conditional probabilities). See Appendix A for a review of formula when class-conditional distributions are Gaussians. Therefore, we need to set good lower and upper bounds to characterize Bayes error B_e (or the probability of error P_e).

A first upper bound on the probability of error P_e is the Bhattacharrya bound [10]:

$$P_e \leqslant \sqrt{w_1 w_2} \times \rho(p_1, p_2), \tag{27}$$

with $\rho(p_1, p_2) = \int \sqrt{p_1(x) p_2(x)} \mathrm{d}x$ denoting the *Bhattacharrya coefficient*. This first upper bound was tightened by Chernoff [4] as follows: We have

$$P_e \leqslant w_1^\alpha w_2^{1-\alpha} \times \rho_\alpha(p_1, p_2), \tag{28}$$

with $\rho_\alpha(p_1, p_2) = \int (p_1(x))^\alpha (p_2(x))^{1-\alpha} \mathrm{d}x$, the *$\alpha$-Chernoff coefficient* defined for $\alpha \in (0, 1)$. Therefore the tightest upper bound is:

$$P_e \leqslant w_1^{\alpha^*} w_2^{1-\alpha^*} \times \rho_*(p_1, p_2), \tag{29}$$

with ρ_* the *Chernoff coefficient* obtained from the following minimization problem:

$$\rho_*(p_1, p_2) = \min_{\alpha \in [0,1]} \rho_\alpha(p_1, p_2), \tag{30}$$

where $\alpha^* \in [0, 1]$ denotes the optimal value that minimizes $\rho_\alpha(p_1, p_2)$. Note that the Bhattacharyya coefficient is a particular case of the Chernoff α-coefficient (obtained for $\alpha = \frac{1}{2}$): $\rho = \rho_{\frac{1}{2}}$.

The Bhattacharyya, α-Chernoff and Chernoff coefficients ρ, ρ_α and $\rho_* = \rho_{\alpha^*}$ can be interpreted a *similarity measures* between distributions defined by measuring the degree of overlap of their densities. Those coefficients are also called affinities.

Remark 1 (*Affinity coefficients and divergences*). The Chernoff α-coefficient ρ_α, the Chernoff coefficient ρ_* and the Bhattacharrya coefficient ρ $\left(\rho = \rho_{\frac{1}{2}}\right)$ provide upper bounds on $P_e : 0 < P_e \leqslant \frac{1}{2}\rho_* \leqslant \frac{1}{2}\rho \leqslant \frac{1}{2}$ (for $w_1 = w_2 = \frac{1}{2}$, see Eqs. (29) and (27)). We can transform any affinity coefficient $0 < A \leqslant 1$ into a corresponding divergence by applying a monotonously increasing function f to $\frac{1}{A}$ (with $\frac{1}{A} \in [1, \infty)$). By choosing $f(x) = \log(x)$, we end-up with the traditional divergences between statistical distributions: α-Chernoff divergence, Chernoff divergence (also called Chernoff information), and Bhattacharrya divergence.

Recall that those upper bounds are useful if they can be computed easily from closed-form formula (which is not the case of B_e nor P_e).

1.4. Closed-form Bhattacharrya/Chernoff coefficients for exponential families

Many usual distributions like Gaussians, Poisson, Dirichlet or Gamma/Beta, etc. distributions are exponential families in disguise [11,16] for which the *skewed affinity coefficient* ρ_α (i.e., similarity distance within $[0, 1]$) can be computed in closed-form. An exponential family is a family $\mathcal{F}$ of distributions:

$$\mathcal{F} = \{p(x; \theta) = \exp(x^\top \theta - F(\theta)) \mid \theta \in \Theta\}, \tag{31}$$

indexed by a parameter $\theta \in \Theta$. Space Θ is the parameter domain, called the natural parameter space [16]. F is a strictly convex and differentiable convex function called the log-normalizer. For example, the multivariate normal (MVN) distributions of mean μ and covariance matrix Σ are exponential families for parameter $\theta = (\theta_1 = \Sigma^{-1}\mu, \theta_2 = \frac{1}{2}\Sigma^{-1}) \in \mathbb{R}^d \times S_d+$ (where S_d^+ denotes the space of symmetric positive definite $d \times d$ matrices). The function F (called log-normalizer) expressed in the natural coordinate system is [15]:

$$F_{\text{MVN}}(\theta_1, \theta_2) = \frac{1}{2}\theta_1^\top \theta_2^{-1} \theta_1 - \log|\theta_2|, \tag{32}$$

where $|\cdot|$ denotes the determinant for a matrix operand.

Wlog., let the class-conditional probabilities p_1 and p_2 belong to the same exponential family, then we have [11,16]:

$$\rho_\alpha(p_1, p_2) = e^{-J_F^{(\alpha)}(\theta_1, \theta_2)}, \tag{33}$$

where $J_F^{(\alpha)}(\theta_1, \theta_2)$ denote the *Jensen skewed divergence* [16]:

$$J_F^{(\alpha)}(\theta_1, \theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2) \geqslant 0. \tag{34}$$

For example, let us consider the multivariate Gaussian family. Then, we get the Chernoff α-coefficient:

$$\rho_\alpha^{MVN}(p_1, p_2) = \frac{|\Sigma_1|^{\frac{\alpha}{2}}|\Sigma_2|^{\frac{1-\alpha}{2}}}{|\alpha\Sigma_1 + (1-\alpha)\Sigma_2|^{\frac{1}{2}}} \exp\left(-\frac{\alpha(1-\alpha)}{2}\Delta\mu^\top(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)\Delta\mu\right), \tag{35}$$

with $\Delta\mu = \mu_2 - \mu_1$. Therefore the Chernoff α-divergence, $D_\alpha^{MVN}(p_1, p_2) = -\log\rho_\alpha^{MVN}(p_1, p_2)$:

$$D_\alpha^{MVN}(p_1, p_2) = \frac{1}{2}\log\frac{|\alpha\Sigma_1 + (1-\alpha)\Sigma_2|}{|\Sigma_1|^\alpha|\Sigma_2|^{1-\alpha}} + \frac{\alpha(1-\alpha)}{2}\Delta\mu^\top(\alpha\Sigma_1 \\ + (1-\alpha)\Sigma_2)\Delta\mu. \tag{36}$$

Setting $\alpha = \frac{1}{2}$, we get the *Bhattacharrya divergence*.

In general, we do not have a closed-form solution for finding the optimal α^* yielding the Chernoff coefficient/information. Nevertheless, the optimal value α^* of α can be exactly characterized [15] using the differential-geometric structure of the statistical manifold of the class-conditional distributions, and yields a fast algorithm to arbitrarily finely approximate Chernoff information ρ_* for members of the same exponential family.

1.5. Outline

This paper is organized as follows: In Section 2 we show how Bayes error is related to the total variation distance. Section 3 presents our generalization of Bhattacharrya and Chernoff upper bounds relying on generalized weighted means. Section 4 illustrates several applications of the technique yielding novel upper bounds for various distributions that do not belong to the exponential families, and Section 4.5 studies the tightness of those bounds. Finally, Section 5 concludes this work. Appendix A recalls the Bayes error formula when class-conditional distributions belong to the univariate or the multivariate Gaussian families.

2. Bayes error and the total variation distance: An identity

Recall Bayes error expression of Eq. (19):

$$B_e = \int p(x)\min(r_1(x), r_2(x))dx \tag{37}$$

$$= \int p(x)\min(c_{11}q_1(x) + c_{12}q_2(x), c_{21}q_1(x) + c_{22}q_2(x))dx, \tag{38}$$

with $q_1(x) = \frac{w_1 p_1(x)}{p(x)}$ and $q_2(x) = \frac{w_2 p_2(x)}{p(x)}$ the *a posteriori* probabilities. Using the mathematical rewriting trick:

$$\min(a, b) = \frac{a+b}{2} - \frac{1}{2}|b - a|, \tag{39}$$

we get:

$$B_e = \frac{1}{2}\int(a_1 p_1(x) + a_2 p_2(x) - |a_2 p_2(x) - a_1 p_1(x)|)dx, \tag{40}$$

where $a_1 = w_1(c_{11} + c_{21})$ and $a_2 = w_2(c_{12} + c_{22})$. Finally, using the fact that $\int p_1(x)dx = \int p_2(x)dx = 1$, we end up with:

$$B_e = \frac{a_1 + a_2}{2} - \text{TV}(a_1 p_1, a_2 p_2), \tag{41}$$

where

$$\text{TV}(p, q) = \frac{1}{2}\int|p(x) - q(x)|dx \tag{42}$$

denotes the *total variation distance* extended to *positive distributions* (i.e., not necessarily normalized probability distributions). In particular, for the probability of error, we have $a_1 = w_1$ and $a_2 = w_2$ (with $a_1 + a_2 = 1$) and get:

$$P_e = \frac{1}{2} - \text{TV}(w_1 p_1, w_2 p_2). \tag{43}$$

Note that $\text{TV}(w_1 p_1, w_2 p_2) = w_1\text{TV}\left(p_1, \frac{w_2}{w_1}p_2\right) = w_2\text{TV}\left(\frac{w_1}{w_2}p_1, p_2\right)$. Therefore in the special case $w_1 = w_2 = \frac{1}{2}$, we get the probability of error related to the total variation distance (a metric) on probability distributions by:

$$P_e = \frac{1}{2}(1 - \text{TV}(p_1, p_2)). \tag{44}$$

For sanity check, notice that when $p_1 = p_2$ (undistinguishable distributions), we have $\text{TV}(p_1, p_2) = 0$ and $P_e = \frac{1}{2}$. Clearly, $0 \leqslant \text{TV}(p_1, p_2) \leqslant 1$ and $0 \leqslant P_e \leqslant \frac{1}{2}$. We summarize the result in the following theorem:

Theorem 1. *The Bayes error B_e for the cost design matrix $C = [c_{ij}]$ is related to the total variation metric distance $\text{TV}(p, q) = \frac{1}{2}\int|p(x) - q(x)|dx$ by $B_e = \frac{a_1 + a_2}{2} - \text{TV}(a_1 p_1, a_2 p_2)$ with $a_1 = w_1(c_{11} + c_{21})$ and $a_2 = w_2(c_{12} + c_{22})$. The identity simplifies for probability of error P_e to $P_e = \frac{1}{2} - \text{TV}(w_1 p_1, w_2 p_2)$.*

Thus if we can compute the total variation distance of class-conditional probabilities p_1 and p_2, we can deduce the probability of error, and vice versa:

$$\text{TV}(p_1, p_2) = 1 - 2P_e \geqslant 0. \tag{45}$$

3. Upper bounds with generalized means

Without loss of generality, consider the probability of error P_e:

$$P_e = \int\min(w_1 p_1(x), w_2 p_2(x))dx = S(w_1 p_1, w_2 p_2). \tag{46}$$

The probability of error can be interpreted as a *similarity measure* $S(w_1 p_1, w_2 p_2)$, extending the definition of *histogram intersection* [20] to continuous domains.

Chernoff [4] made use of the following mathematical trick:

$$\min(a, b) \leqslant a^\alpha b^{1-\alpha}, \quad \forall a, \ b > 0, \tag{47}$$

to define the *Chernoff information* upper bounding P_e:

$$P_e \leqslant w_1^{\alpha^*} w_2^{1-\alpha^*}\rho_*(p_1, p_2), \tag{48}$$

with

$$\rho_*(p_1, p_2) = \min_{\alpha \in [0,1]}\rho_\alpha(p_1, p_2), \quad \rho_\alpha(p_1, p_2) = \int(p_1(x))^\alpha(p_2(x))^{1-\alpha}dx.$$

We shall revisit this technique using the wider scope of *generalized weighted means*.

By definition, a mean $M(a, b)$ is a *smooth* function such that $\min(a, b) \leqslant M(a, b) \leqslant \max(a, b)$. Similarly, we can define a weighted mean as a smooth function $M(a, b; \alpha)$ that fulfills the interness property $M(a, b; \alpha) \in [\min(a, b), \max(a, b)] \ \forall\alpha \in [0, 1]$. Let us consider the *quasi-arithmetic means* (also called Kolmogorov–Nagumo *f*-means [12,14]) for a strictly monotonous generator function *f*:

Lemma 1 [1]. *The quasi-arithmetic weighted mean $M_f(a, b; \alpha) = f^{-1}(\alpha f(a) + (1-\alpha)f(b))$ of two real values a and b for a strictly monotonic function f satisfies the interness property: $\min(a, b) \leqslant M_f(a, b; \alpha) \leqslant \max(a, b)$.*

Proof. Assume f is strictly increasing and $a \leqslant b$, then $f(a) \leqslant f(b)$ and $f(a) \leqslant \alpha f(a) + (1-\alpha)f(b) \leqslant f(b)$. Thus $a \leqslant M_f(a,b;\alpha) \leqslant b$. If $b \leqslant a$, we similarly have $b \leqslant M_f(a,b;\alpha) \leqslant a$. Therefore $\min(a,b) \leqslant M_f(a,b;\alpha) \leqslant \max(a,b)$. The proof is identical when f is strictly decreasing. □

Interestingly, one important property of quasi-arithmetic means is their *dominance relationship*. That is, if $f(x) \leqslant g(x)$ then $M_f(a,b;\alpha) \leqslant M_g(a,b;\alpha)$ (with equality when $a=b$). This property generalizes the well known arithmetic–geometric–harmonic (AGH) inequality property of Pythagorean means:

$$M_{f_a}(a,b;\alpha) \geqslant M_{f_g}(a,b;\alpha) \geqslant M_{f_h}(a,b;\alpha), \tag{49}$$

with $f_a(x) = x, f_g(x) = \log x$ and $h(x) = 1/x$ denoting the generators for the arithmetic, geometric, and harmonic means, respectively. That is, we have:

$$\alpha a + (1-\alpha)b \geqslant a^{\alpha}b^{1-\alpha} \geqslant \frac{ab}{\alpha a + (1-\alpha)b}. \tag{50}$$

Similarly to Chernoff [4], we define the following *generalized affinity coefficient* ρ_f using generalized weighted means as follows:

Definition 1. The Chernoff-type similarity coefficient (affinity) for a strictly monotonous function f is defined by:

$$\rho_*^f(p_1, p_2) = \min_{\alpha \in [0,1]} \int M_f(p_1(x), p_2(x); \alpha) dx \leqslant \int p_1(x) dx = 1 \tag{51}$$

and define the generalized Chernoff information as:

Definition 2. The Chernoff-type information for a strictly monotonous function f is defined by:

$$C_f(p_1, p_2) = -\log \rho_*^f(p_1, p_2)$$
$$= \max_{\alpha \in [0,1]} -\log \int M_f(p_1(x), p_2(x); \alpha) dx \geqslant 0. \tag{52}$$

Corollary 1. *The traditional Chernoff similarity, information, and upper bound are obtained by choosing the weighted geometric mean, by setting the generator* $f_{\text{Chernoff}}(x) = \log(x)$ *(with* $f_{\text{Chernoff}}^{-1}(x) = \exp(x)$*). We get* $M_{f_{\text{Chernoff}}}(p_1(x), p_2(x); \alpha) = p_1(x)^{\alpha}p_2(x)^{1-\alpha}$.

When we do not optimize over the parameter α, but assume it fixed to $\frac{1}{2}$, we extend the Bhattacharyya coefficient and divergence as follows:

Definition 3. The generalized skew Bhattacharyya-type similarity coefficient (affinity) for a strictly monotonous function f is defined by:

$$\rho_\alpha^f(p_1, p_2) = \int M_f(p_1(x), p_2(x); \alpha) dx \leqslant \int p_1(x) dx = 1 \tag{53}$$

and the generalized skew Bhattacharyya-type divergence is defined as $B_\alpha^f = -\log \rho_\alpha^f(p_1, p_2)$. The generalized Bhattacharyya coefficient $\rho^f(p_1, p_2) = \int M_f(p_1(x), p_2(x); \frac{1}{2}) dx$ and divergence $B^f(p_1, p_2) = -\log \rho^f(p_1, p_2)$.

Theorem 2. *Using quasi-arithmetic means, we can bound the probability of error as follows:*

$$P_e = \int \min(w_1 p_1(x), w_2 p_2(x)) dx$$
$$\leqslant \int M_f(w_1 p_1(x), w_2 p_2(x); \alpha) dx. \tag{54}$$

The upper bound proves useful for well-chosen f yielding closed-form expression of the rhs.

In particular, by choosing the power means M_{f_β} obtained for $f_\beta(x) = x^\beta$, we get a *tight bound* in the limit case since $M_{f_\beta}(p, q) \to \min(p, q)$ when $\beta \to -\infty$. However, in order for the generalized affinity, distance and upper bound to be useful, we need to be able to compute them in *closed-form* for some statistical distribution families. We illustrate how to derive closed form formula for ρ_f and closed form upper bounds on P_e for several statistical distribution families.

The generalized Bhattacharyya upper bound $\rho_{\frac{1}{2}}^f$ is obtained by setting $\alpha = \frac{1}{2}$:

$$P_e = \int \min(w_1 p_1(x), w_2 p_2(x)) dx \leqslant B_f(w_1 p_1(x), w_2 p_2(x)), \tag{55}$$

$$B_f(w_1 p_1(x), w_2 p_2(x)) = \int M_f\left(w_1 p_1(x), w_2 p_2(x); \frac{1}{2}\right) dx. \tag{56}$$

In order for the Chernoff information to improve over the Bhattacharyya bound, we need the quasi-arithmetic α-weighted mean to be a *convex function* with respect to parameter α. For the geometric mean, we check that:

$$M_{f_g}(a,b;\alpha) = e^{\alpha \log \frac{a}{b} + \log b}, \tag{57}$$

is strictly convex with respect to α (since $\frac{d^2}{d\alpha^2} M_{f_g}(a,b;\alpha) = (\log \frac{a}{b})^2 M_{f_g}(a,b;\alpha) > 0$ for $a \neq b$). Similarly, the weighted harmonic mean is convex with respect to α since $\frac{d^2}{d\alpha^2} M_{f_h}(a,b;\alpha) = 2ab(a-b)^2(\alpha(a-b)+b)^{-3} > 0$.

Remark 2. Note that not all quasi-arithmetic means yield convex weighted means. Indeed, let $M_f'(a,b;\alpha) = f(\alpha(f^{-1}(a) - f^{-1}(b)) + f^{-1}(b))$ then $\frac{d^2}{d\alpha^2} M_f'(a,b;\alpha) = (f^{-1}(a) - f^{-1}(b))^2 f''(\alpha(f^{-1}(a) - f^{-1}(b)) + f^{-1}(b))$. Functions f and f^{-1} are strictly monotonous but can be convex, concave, or arbitrary in general.

Remark 3. Sometimes, we prefer to parameterize the weighted mean as the smooth interpolant from a to b, when α varies from 0 to 1 (kind of geodesic parameterization). In that case, we may prefer the parameterization $M_f(a,b;\alpha') = f^{-1}((1-\alpha')f(a) + \alpha' f(b))$. This is not important for Bhattacharyya-type symmetric bounds nor for Chernoff-type bounds that optimize over the α range (or equivalently over the α' range).

Let us now examine how the "quasi-arithmetic bounding techniques" apply for several families of statistical distributions.

4. Some illustrating examples

4.1. Geometric means and the Chernoff bound for exponential families

First, we recall the well-known formula [11,16] for the case of exponential families.

In order to compute a closed form for the right-hand side of Eq. (58):

$$P_e \leqslant \int M_f(w_1 p_1(x), w_2 p_2(x); \alpha) dx, \tag{58}$$

$$\leqslant \int f^{-1}(\alpha f(w_1 p_1) + (1-\alpha)f(w_2 p_2)) dx, \tag{59}$$

we consider the *geometric mean* obtained for $f(x) = \log x$. Since $p_1(x) = \exp(x^\top \theta_1 - F(\theta_1))$ and $p_2(x) = \exp(x^\top \theta_2 - F(\theta_2))$ belong to the exponential families, we get:

$$M_f(w_1 p_1(x), w_2 p_2(x); \alpha) = e^{\alpha \log w_1 p_1(x) + (1-\alpha) \log w_2 p_2(x)}, \tag{60}$$

$$= w_1^\alpha w_2^{1-\alpha} p_1^\alpha(x) p_2^{1-\alpha}(x). \tag{61}$$

It follows that:

$$P_e \leqslant w_1^\alpha w_2^{1-\alpha} \int f^{-1}(\alpha f(p_1(x)) + (1-\alpha)f(p_2(x)))\mathrm{d}x. \quad (62)$$

Remark 4. In fact, the geometric mean is a limit case of a family of linear-scale free means defined for $f_\alpha(x) = x^{\frac{1-\alpha}{2}}$ with $f_1(x) = \log x$. In general, in order to slide the *a priori* weights w_1 and w_2 out of the integral, we would like to use a homogeneous function f (with $f(\lambda x) = g(\lambda) f(x)$).

Furthermore, since $m_\alpha(x; \theta_1, \theta_2) = \alpha \log(p_1(x)) + (1-\alpha) \log(p_2(x)) = x^\top(\alpha\theta_1 + (1-\alpha)\theta_2) - \alpha F(\theta_1) - (1-\alpha)F(\theta_2)$, we would like to get $f^{-1}(m_\alpha(x; \theta_1, \theta_2))$ as $c_{\theta_1, \theta_2; \alpha} p(x; \theta_{12}^{(\alpha)})$ so that we can slide the integral operand inside the expression, and use the fact that we recognize a member $\theta_{12}^{(\alpha)}$ of the exponential family so that its integration over the support is 1. For members of the *same* exponential family, we have:

$$f^{-1}(m_\alpha(x; \theta_1, \theta_2)) = e^{F(\alpha\theta_1 + (1-\alpha)\theta_2) - \alpha F(\theta_1) - (1-\alpha)F(\theta_2)}$$
$$\times p(x; \alpha\theta_1 + (1-\alpha)\theta_2), \quad (63)$$
$$= e^{-J_F^{(\alpha)}(\theta_1, \theta_2)} p(x; \underbrace{\alpha\theta_1 + (1-\alpha)\theta_2}_{\theta_{12}^{(\alpha)}}) \quad (64)$$

Thus

$$P_e \leqslant w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1, \theta_2)} \int p(x; \alpha\theta_1 + (1-\alpha)\theta_2)\mathrm{d}x. \quad (65)$$

Since the natural parameter space Θ is convex for exponential families, we have $\theta_{12}^{(\alpha)} = \alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta$ and therefore $\int p(x; \alpha\theta_1 + (1-\alpha)\theta_2)\mathrm{d}x = 1$. We end up with:

$$P_e \leqslant \min_{\alpha \in [0,1]} w_1^\alpha w_2^{1-\alpha} e^{-J_F^{(\alpha)}(\theta_1, \theta_2)}. \quad (66)$$

Definition 4. The α-Chernoff distance is an asymmetric statistical distance defined by $\rho_\alpha(p_1, p_2) = -\log p_1^\alpha(x) p_2^{1-\alpha}(x)\mathrm{d}x \geqslant 0$. The Bhattacharrya symmetric distance $\rho_{\frac{1}{2}}(p_1, p_2)$ is a particular member of the family of α-Chernoff distances.

Thus we always have $\rho_{\frac{1}{2}}(p_1, p_2) \geqslant \rho_*(p_1, p_2) = \min_{\alpha \in [0,1]} \rho_\alpha(p_1, p_2)$.

The optimal Chernoff bound is obtained for the optimized weight α^* that has been characterized geometrically on the statistical manifold [15].

To derive other Chernoff-type upper bounds, we shall therefore consider non-exponential families of distributions. The most prominent family, to start with, is the Cauchy family (a member of the Student *t*-distribution families) and the multivariate Pearson type VII elliptical distributions [19] (with its scaled multivariate *t*-distributions).

4.2. Harmonic means and the Chernoff-type bound for Cauchy distributions

Consider the family of Cauchy distributions with density:

$$p(x; s) = \frac{1}{\pi} \frac{s}{x^2 + s^2}, \quad (67)$$

defined over the support $\mathbb{R}$. This is a *scale family* with heavy tails indexed by a scale parameter s:

$$p(x; s) = \frac{1}{s} p_0\left(\frac{x}{s}\right), \quad p_0(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}, \quad (68)$$

where $p_0(x)$ denotes the standard Cauchy distribution C_0. Cauchy distributions do not belong to the exponential families. (Indeed, the mean is undefined.) Let us take the harmonic mean $M_H = M_f$ defined for the strictly monotonous generator $f(x) = f^{-1}(x) = \frac{1}{x}$. We have:

$$P_e = \int \min(w_1 p_1(x), w_2 p_2(x))\mathrm{d}x$$
$$\leqslant M_f(w_1 p_1(x), w_2 p_2(x); \alpha)\mathrm{d}x. \quad (69)$$

Wlog., to simplify calculations exhibiting the method, consider $w_1 = w_2 = \frac{1}{2}$.

$$P_e \leqslant \int M_H\left(\frac{1}{2}p_1(x), \frac{1}{2}p_2(x); \alpha\right)\mathrm{d}x, \quad (70)$$
$$\leqslant \frac{1}{2} \int \frac{p_1(x)p_2(x)}{(1-\alpha)p_1(x) + \alpha p_2(x)}\mathrm{d}x, \quad (71)$$
$$\leqslant \frac{1}{2} \int \frac{\frac{s_1}{\pi(x^2 + s_1^2)} \frac{s_2}{\pi(x^2 + s_2^2)}}{(1-\alpha)\frac{s_1}{\pi(x^2 + s_1^2)} + \alpha\frac{s_2}{\pi(x^2 + s_2^2)}}\mathrm{d}x, \quad (72)$$
$$\leqslant \frac{1}{2} \int \frac{s_1 s_2}{\pi((1-\alpha)s_1(x^2 + s_2^2) + \alpha s_2(x^2 + s_1^2))}\mathrm{d}x, \quad (73)$$
$$\leqslant \frac{1}{2} \int \frac{s_1 s_2}{\pi(((1-\alpha)s_1 + \alpha s_2)x^2 + (1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2)}\mathrm{d}x, \quad (74)$$
$$\leqslant \frac{1}{2} \frac{s_1 s_2}{((1-\alpha)s_1 + \alpha s_2)s_\alpha} \underbrace{\int \frac{1}{\pi} \frac{s_\alpha}{x^2 + s_\alpha^2}\mathrm{d}x}_{=1}, \quad (75)$$

since $s_\alpha > 0$ belongs to the parameter space Θ, with:

$$s_\alpha = \sqrt{\frac{(1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2}{(1-\alpha)s_1 + \alpha s_2}}. \quad (76)$$

Thus the weighted harmonic mean provides a Chernoff-type upper bound:

$$P_e \leqslant \frac{1}{2} \frac{s_1 s_2}{((1-\alpha)s_1 + \alpha s_2)\sqrt{\frac{(1-\alpha)s_1 s_2^2 + \alpha s_2 s_1^2}{(1-\alpha)s_1 + \alpha s_2}}}. \quad (77)$$

A sanity check $s_1 = s_2 = s$ shows that $P_e = \frac{1}{2}$, as expected (class-conditional distributions are not distinguishable).

The Bhattacharrya-type bound obtained for $\alpha = \frac{1}{2}$ yields:

$$s_{\frac{1}{2}} = \sqrt{\frac{s_1 s_2^2 + s_2 s_1^2}{s_1 + s_2}}, \quad (78)$$

and the upper bound:

$$P_e \leqslant \frac{s_1 s_2}{\sqrt{(s_1 + s_2)(s_1 s_2^2 + s_2 s_1^2)}}. \quad (79)$$

The harmonic mean $M_H(a, b; \alpha)$ is *linear-scale free*: $M_H(\lambda a, \lambda b; \alpha) = \lambda M_H(a, b; \alpha)$. Let $\lambda = \frac{s_2}{s_1}$. Then we write the probability of error as:

$$P_e \leqslant \frac{1}{2} \frac{\lambda}{\sqrt{(1-\alpha + \alpha\lambda)((1-\alpha)\lambda^2 + \alpha\lambda)}} = P_e^{(\alpha)}. \quad (80)$$

In particular, we upper bound the probability of error by the following Bhattacharyya bound:

$$P_e \leqslant P_e^{(\frac{1}{2})} = \frac{\sqrt{\lambda}}{\lambda + 1}. \quad (81)$$

The Chernoff-type bound proceeds by minimizing Eq. (80) with respect to α. We have $P_e^{(0)} = P_e^{(1)} = \frac{1}{2}$. To find the minimum value over the α-range $[0, 1]$, we study function $P_e^{(\alpha)}$. Using a computer-algebra system,[3] we find that it is a convex function that always admits a minimum at $\alpha = \frac{1}{2}$. Indeed, the derivative of P_e with respect to α is:

---

[3] Namely, Wolfram Alpha online, http://www.wolframalpha.com.

$$\frac{d}{d\alpha} P_e(\alpha; \lambda) = \frac{\frac{1}{4}(\lambda - 1)^2 \lambda^2 (2\alpha - 1)}{(\lambda((\lambda - 1)\alpha + 1)(-\alpha\lambda + \lambda + \alpha))^{3/2}}, \tag{82}$$

that is zero if and only if $\alpha = \frac{1}{2}$. This is a remarkable example that shows that the Chernoff bound amounts to the Bhatthacharrya bound.

Let us compute the total variation distance between two scaled Cauchy distributions $a_1 p(x; s_1)$ and $a_2 p(x; s_2)$ defined over the real-line support $\mathbb{R}$ with:

$$p(x; s) = \frac{1}{\pi} \frac{s}{x^2 + s^2} \tag{83}$$

For $s_1 \neq s_2$, the two distinct positive densities intersect in exactly two values of the support:

$$x_1 = -\frac{\sqrt{s_1 s_2 (a_2 s_1 - a_1 s_2)}}{\sqrt{a_1 s_1 - a_2 s_2}}, \tag{84}$$

$$x_2 = \frac{\sqrt{s_1 s_2 (a_2 s_1 - a_1 s_2)}}{\sqrt{a_1 s_1 - a_2 s_2}}. \tag{85}$$

In particular, when $a_1 = a_2 = 1$ (i.e., probability densities), we have $x_1 = -\sqrt{s_1 s_2}$ and $x_2 = \sqrt{s_1 s_2}$. By abuse of notations, we let $x_0 = -\infty$ and $x_3 = \infty$, and apply the generic 1D total variation formula of Eq. (114) with $k = 2$:

$$\text{TV}(a_1 p_1, a_2 p_2) = \frac{1}{2} \sum_{i=1}^{k+1} |(P_1(x_i) + P_2(x_{i+1}) - P_1(x_{i-1}) - P_2(x_i))|. \tag{86}$$

Since the Cauchy scaled cumulative distribution is $P_i(x) = a_i \left(\frac{1}{\pi} \arctan \left(\frac{x}{s_i} \right) + \frac{1}{2} \right)$ with $P_i(x_0) = 0$ and $P_i(x_3) = a_i$.

In particular, the probability of error when $w_1 = w_2 = \frac{1}{2}$ is:

$$P_e = \frac{1}{2}(1 - \text{TV}(p_1, p_2)), \tag{87}$$

$$\text{TV}(p_1, p_2) = |P_1(\sqrt{s_1 s_2}) - P_1(-\sqrt{s_1 s_2}) - P_2(\sqrt{s_1 s_2}) + P_2(-\sqrt{s_1 s_2})|. \tag{88}$$

That is, we find an exact analytic expression for the total variation (and hence for Bayes error B_e and the probability of error P_e):

$$\text{TV}(p_1, p_2) = \frac{1}{\pi} \left(\arctan \left(\sqrt{\frac{s_2}{s_1}} \right) - \arctan \left(-\sqrt{\frac{s_2}{s_1}} \right) + \arctan \left(-\sqrt{\frac{s_1}{s_2}} \right) - \arctan \left(\sqrt{\frac{s_1}{s_2}} \right) \right). \tag{89}$$

Using the identity $\arctan(x) + \arctan(1/x) = \frac{\pi}{2}$ and the fact that $\arctan(-x) = -\arctan(x)$, we get a more compact formula:

$$\text{TV}(p_1, p_2) = \frac{2}{\pi} \left(\arctan \left(\sqrt{\frac{s_2}{s_1}} \right) - \arctan \left(\sqrt{\frac{s_1}{s_2}} \right) \right). \tag{90}$$

Remark 5. Note that for Cauchy distributions with scale parameter s_1 and s_2, we have $\text{TV}(s_1, s_2) = \text{TV}(\lambda s_1, \lambda s_2), \forall \lambda > 0$ (because $\sqrt{\frac{\lambda s_1}{\lambda s_2}} = \sqrt{\frac{s_1}{s_2}}$). Therefore, we may renormalize by considering $s_1 \leftarrow 1$ and $s_2 \leftarrow \frac{s_2}{s_1}$.

It follows that the probability of error is:

$$P_e = \frac{1}{2} - \frac{1}{\pi} \left(\arctan \left(\sqrt{\lambda} \right) - \arctan \left(\sqrt{1/\lambda} \right) \right), \tag{91}$$

$$= 1 - \frac{2}{\pi} \arctan \left(\sqrt{\lambda} \right), \quad \lambda = \frac{s_2}{s_1}. \tag{92}$$

Consider the following numerical example: $s_1 = 10$ and $s_2 = 50$ ($w_1 = w_2 = \frac{1}{2}$). Then, we get $P_e \sim 0.2677$, the Bhattacharyya-type bound $B \sim 0.3726$ and the Chernoff-type bound $C = B$.

Remark 6. Let us study the tightness of the upper bound. We can express analytically the gap as $\Delta = C - P_e = C - \frac{1}{2} + \text{TV}(p_1, p_2)$. That is, we get:

$$\Delta(\lambda) = \frac{\sqrt{\lambda}}{1 + \lambda} - 1 + \frac{2}{\pi} \arctan \left(\sqrt{\lambda} \right) > 0. \tag{93}$$

Note that $\lambda = 1$, we have $\Delta(1) = 0$: That is, the gap is tight when distributions coincide. Using a computer-algebra system, we find that the gap is maximized at $\lambda = \frac{2 + \pi}{\pi - 2} \sim 4.5$ and $\Delta_{\max} \sim 0.1$.

This Cauchy case study illustrates well that the geometric mean may not always be the most appropriate weighted mean to use to upper bound P_e. Note that for Cauchy distributions, we also obtained an analytic form of P_e and B_e using the total variation distance expressed using the cumulative distribution.

We showed that the harmonic mean is tailored to derive closed-form bound for the Cauchy distributions. However, we may apply the harmonic mean to other distributions. This has in fact be done in the literature detailed in the following remark:

Remark 7. The harmonic bean $M_H(a, b) = \frac{2ab}{(a+b)}$ has been used for defining the *nearest neighbor error bound* [7], always better than the Bhattacharrya bound (obtained for $\alpha = \frac{1}{2}$).

(Note that for exponential families, the nearest neighbor error bound is not available in closed-form.)

4.3. Pearson type VII distributions

Consider the d-dimensional elliptically symmetric Pearson type VII distribution [19,2]) with density:

$$p(x; \mu, \Sigma, \lambda) = \pi^{-\frac{d}{2}} \frac{\Gamma(\lambda)}{\Gamma \left(\lambda - \frac{d}{2} \right)} |\Sigma|^{-\frac{1}{2}} \left(1 + (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)^{-\lambda}, \tag{94}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ denotes the Gamma function extending the factorial function (i.e., $\Gamma(1) = 1, \Gamma(x + 1) = x\Gamma(x)$ and $\Gamma(n) = (n - 1)!$ for $n \in \mathbb{N}$). Parameter $\lambda > \frac{d}{2}$ represents the degree of freedom [19].

For sake of simplicity, wlog., let us consider the zero-centered sub-family with $\mu = 0, \nu = 1$ and $\lambda > \frac{d}{2}$ fixed. This sub-family is defined on the cone of symmetric positive definite matrices $\Theta = \{\Sigma \mid \Sigma \succ 0\}$ with density:

$$p(x; \Sigma) = c_d(\lambda) |\Sigma|^{-\frac{1}{2}} (1 + x^\top \Sigma^{-1} x)^{-\lambda}, \tag{95}$$

where $c_d(\lambda) = \pi^{-\frac{d}{2}} \frac{\Gamma(\lambda)}{\Gamma \left(\lambda - \frac{d}{2} \right)}$ is the normalizing constant [13]. This multivariate family does not belong to the exponential families. Consider the α-weighted f-mean with $f(x) = x^{-\frac{1}{\lambda}}$, for prescribed $\lambda > \frac{d}{2}$ (and $f^{-1}(x) = x^{-\lambda}$).

We have:

$$\alpha f(p_1) = \alpha c_d(\lambda)^{-1/\lambda} |\Sigma_1|^{\frac{1}{2\lambda}} (1 + x^\top \Sigma_1^{-1} x), \tag{96}$$

$$(1 - \alpha) f(p_2) = (1 - \alpha) c_d(\lambda)^{-1/\lambda} |\Sigma_2|^{\frac{1}{2\lambda}} (1 + x^\top \Sigma_1^{-2} x). \tag{97}$$

Let $c_1 = c_d(\lambda)^{-1/\lambda} |\Sigma_1|^{\frac{1}{2\lambda}}$ and $c_2 = c_d(\lambda)^{-1/\lambda} |\Sigma_2|^{\frac{1}{2\lambda}}$. Denote by:

$$c_\alpha = \alpha c_1 + (1 - \alpha) c_2, \tag{98}$$

$$= c_d(\lambda)^{-1/\lambda} (\alpha |\Sigma_1|^{\frac{1}{2\lambda}} + (1 - \alpha) |\Sigma_2|^{\frac{1}{2\lambda}}), \tag{99}$$

we get:

$$\alpha f(p_1) + (1 - \alpha) f(p_2) = \alpha c_1 (1 + x^\top \Sigma_1^{-1} x) + (1 - \alpha) c_2 (1 + x^\top \Sigma_2^{-1} x), \tag{100}$$

$$= c_\alpha (1 + x^\top \Sigma_\alpha^{-1} x), \tag{101}$$

with

$$\Sigma_\alpha^{-1} = \frac{\alpha c_1 \Sigma_1^{-1} + (1-\alpha)c_2 \Sigma_2^{-1}}{c_\alpha}, \tag{102}$$

$$= \frac{\alpha |\Sigma_1|^{\frac{1}{2\lambda}}\Sigma_1^{-1} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}}\Sigma_2^{-1}}{(\alpha|\Sigma_1|^{\frac{1}{2\lambda}} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}})} \succ 0. \tag{103}$$

Therefore,

$$f^{-1}(\alpha f(p_1) + (1-\alpha)f(p_2)) = c_\alpha^{-\lambda}(1 + x^\top \Sigma_\alpha^{-1} x)^{-\lambda}, \tag{104}$$

$$= c_\alpha^{-\lambda}|\Sigma_\alpha|^{\frac{1}{2}}\frac{1}{c_d(\lambda)}p(x;\Sigma_\alpha). \tag{105}$$

It follows that:

$$P_e \leqslant \frac{1}{2}(\alpha|\Sigma_1|^{\frac{1}{2\lambda}} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}})^{-\lambda}|\Sigma_\alpha|^{\frac{1}{2}}\underbrace{\int p(x;\Sigma_\alpha)\mathrm{d}x}_{=1}, \tag{106}$$

$$= \frac{1}{2}(\alpha|\Sigma_1|^{\frac{1}{2\lambda}} + (1-\alpha)|\Sigma_2|^{\frac{1}{2\lambda}})^{-\lambda}|\Sigma_\alpha|^{\frac{1}{2}}, \tag{107}$$

since $\Sigma_\alpha \in \Theta$.

The Pearson type VI distribution is related to the multivariate t-distributions [13].

4.4. Central multivariate t-distributions

The multivariate t-distribution (MVT, centered at $\mu = 0$) with $v \geqslant 1$ *degrees of freedom* is defined for a positive definite matrix $\Sigma \succ 0$ (the *scale matrix*) by the following density:

$$p(x;\Sigma) = c_{d,v}|\Sigma|^{-\frac{1}{2}}\left(1 + \frac{1}{v}x^\top \Sigma^{-1}x\right)^{-\frac{v+d}{2}}, \tag{108}$$

where $c_{d,v} = \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})(v\pi)^{\frac{d}{2}}}$ is the constant normalizing the distribution. The covariance matrix is $\frac{v}{v-2}\Sigma$.

Let $t = -\frac{v+d}{2}$, and consider $f(x) = x^{\frac{1}{t}}$, with functional inverse $f^{-1}(x) = x^t$. Using a technique similar to the Pearson bound, after massaging the mathematics, we find that (for $w_1 = w_2 = \frac{1}{2}$) $P_e \leqslant \frac{1}{2}\rho_\alpha^{\mathrm{MVT}}(\Sigma_1,\Sigma_2)$ (for $\alpha \in [0,1]$), with:

$$\rho_\alpha^{\mathrm{MVT}}(\Sigma_1,\Sigma_2) = (\alpha|\Sigma_1|^{-\frac{1}{2t}} + (1-\alpha)|\Sigma_2|^{-\frac{1}{2t}})^t|\Sigma_\alpha'|^{\frac{1}{2}}, \tag{109}$$

and

$$\Sigma_\alpha' = \left(\frac{\alpha|\Sigma_1|^{-\frac{1}{2t}}\Sigma_1^{-1} + (1-\alpha)|\Sigma_2|^{-\frac{1}{2t}}\Sigma_2^{-1}}{\alpha|\Sigma_1|^{-\frac{1}{2t}} + (1-\alpha)|\Sigma_2|^{-\frac{1}{2t}}}\right)^{-1}. \tag{110}$$

Note that when $v \to \infty$, and $t = -\frac{2}{v+d} \to 0$, the multivariate t-distribution (MVT) tend to a multivariate Normal distribution (MVN) with covariance matrix Σ. The power mean induced by $f(x) = x^{\frac{1}{t}}$ (with $t = -\frac{v+d}{2}$) tends to the geometric mean, and we get the well-known Bhattacharyya coefficient bound (for $\alpha = \frac{1}{2}$) on central multivariate Gaussians [16] (see Eq. (35)):

$$P_e \leqslant \frac{1}{2}\rho^{\mathrm{MVN}}(\Sigma_1,\Sigma_2), \quad \rho^{\mathrm{MVN}}(\Sigma_1,\Sigma_2) = \frac{|\Sigma_1|^{\frac{1}{4}}|\Sigma_2|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{\frac{1}{2}}}. \tag{111}$$

4.5. Assessing the Bhattacharyya-type and the Chernoff-type upper bounds

Let us recall the inequality on the probability of error P_e between two distributions p_1 and p_2 with equal prior ($w_1 = w_2 = \frac{1}{2}$):

$$P_e(p_1,p_2) = \frac{1}{2}(1 - \mathrm{TV}(p_1,p_2)) \leqslant \frac{1}{2}\rho_*^f(p_1,p_2) \leqslant \frac{1}{2}\rho^f(p_1,p_2) \leqslant \frac{1}{2}. \tag{112}$$

The left hand side has been elucidated in Eq. (44) and the right hand side is the Chernoff-type/Bhattacharyya-type similarity coefficients which should be available in closed-form for fast calculation. We are interested in characterizing the gaps $\Delta = \rho(p_1,p_2) - P_e$ and $\Delta_* = \rho_*(p_1,p_2) - P_e$ between the Bhattacharyya/Chernoff upper bounds and P_e (with $\Delta_* \leqslant \Delta$). We start with some simple cases of univariate distributions, where P_e can be expressed analytically, and then considered the multivariate distributions where P_e need to be stochastically estimated.

4.5.1. Simple cases: univariate distributions

For univariate densities, we may calculate the total variation by computing the roots of $p_1(x) = p_2(x)$ and then using the cumulative distributions $P(t) = \int p(x \leqslant t)\mathrm{d}x$ to explicit a formula. Assume $x_1,\ldots,x_k$ are the k roots, and by abuse of notations, let $x_0 = x_{\min}$ and $x_{k+1} = x_{\max}$ denote the extra endpoints of the distribution support ($x \in [x_{\min},x_{\max}]$). When the support is the full real line ($\mathrm{supp}(p_i) = \mathbb{R}$), we set $x_{\min} = -\infty$ and $x_{\max} = \infty$. We have:

$$\mathrm{TV}(p_1,p_2) = \frac{1}{2}\sum_{i=1}^{k+1}\left|\int_{x_{i-1}}^{x_i}(p_1(x) - p_2(x))\mathrm{d}x\right|, \tag{113}$$

$$= \frac{1}{2}\left|\sum_{i=1}^{k+1}(P_1(x_i) + P_2(x_{i+1}) - P_1(x_{i-1}) - P_2(x_i))\right|. \tag{114}$$

This scheme allows one to compute the total variation distance of many *univariate* distributions like the Gaussian (see Eq. (A.13)), Rayleigh, Cauchy, etc. distributions. Therefore, we get analytic expressions of Bayes error relying on the *cumulative distribution function* (CDF) for many univariate distributions.

Those the Bhattacharyya gaps for the Cauchy or 1D Gaussians can be written analytically. Section 4.2 already addressed the gap for the Cauchy distributions. Next, we consider the general multivariate case (that includes those univariate examples) and report on our numerical experiments.

4.5.2. The case of multivariate distributions

We consider multivariate t-distributions [8,9] (MVT) that includes the multivariate normal (MVN) distributions in the limit case (when the number of degree of freedom tends to infinity). To perform experiments, we used the `mvtnorm` package[4] on the R software platform.[5]

Since the total variation (TV) does not admit a closed-form formula (nor the probability of error P_e), we estimate those quantities by performing stochastic integrations as follows:

$$\widehat{P_e}(p_1,p_2) = \frac{1}{2}(1 - \widehat{\mathrm{TV}}(p_1,p_2)), \tag{115}$$

$$\widehat{\mathrm{TV}}(p_1,p_2) = \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{p_1(x)}|p_1(x) - p_2(x)|, \tag{116}$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\left|1 - \frac{p_2(x_i)}{p_1(x_i)}\right|, \tag{117}$$

where $x_1,\ldots,x_n$ are n identically and independently variates of p_1. Stochastic integration guarantees convergence to the true value in the limit: $\lim_{n\to\infty}\widehat{P_e}(p_1,p_2) = P_e(p_1,p_2)$.

To give a numerical example, consider central bidimensional t-distributions with $v = 6$ and scale matrices $\Sigma_1 = I$ and $\Sigma_2 = 10I$, where I denotes the identity matrix. Running the `mvtTotalVariation (df,sigma1,sigma2,n)` code,[6] we get the following

---

Table 1
Experimental results for central multivariate t-distributions (MVT) with $\nu = 6, \Sigma_1 = I$ and $\Sigma_2 = \lambda I$. We have: $\widehat{P_e} \leqslant \frac{1}{2}\rho_*^{\text{MVT}} \leqslant \frac{1}{2}\rho^{\text{MVT}}$ where ρ_*^{MVT} and ρ^{MVT} denotes the Chernoff and Bhattacharyya coefficient, respectively. Observe that the Chernoff upperbound is tighter (but of the same order) than the Bhattacharyya bound (for $\alpha^* \neq \frac{1}{2}$). Those upper bounds improve over the naive $\frac{1}{2}$ bound.

Dimension	λ	$\widehat{P_e}$	ρ_*^{MVT}	$\widehat{\alpha^*}$	ρ^{MVT}
d = 2	$\lambda = 3$	0.3302	0.4471817	0.455	0.4475742
d = 3	$\lambda = 4$	0.2578	0.3951277	0.462	0.3956298
d = 5	$\lambda = 6$	0.16215	0.2943599	0.487	0.2944589
d = 10	$\lambda = 11$	0.06045	0.1400655	0.548	0.141438
d = 15	$\lambda = 16$	0.02845	0.07442729	0.592	0.07841622
d = 20	$\lambda = 21$	0.0167	0.04396945	0.625	0.04945252

estimates for $\widehat{TV}$: 0.3210709 ($n = 100$), 0.3479519 ($n = 1000$), 0.3472926 ($n = 10{,}000$), 0.347538 ($n = 100{,}000$). We chose $n = 10{,}000$ in the following experiments.

Consider central t-distributions ($\mu = 0$). We implemented the closed-form formula of Eq. (109) to calculate the α-Chernoff coefficient $\rho_\alpha^{\text{MVT}}(\Sigma_1, \Sigma_2)$. The optimal Chernoff coefficient (and the exponent α^*) is approximated by discretizing into 1000 steps the unit range for α. We consider $\Sigma_1 = I$ and $\Sigma_2 = \lambda I$ (with $\nu = 6$) for $\lambda = d + 1$ and various values of the dimension. (Indeed, after an appropriate "whitening" transformation [7], we may assume wlog. that one parameter matrix is the identity while the other is diagonal.) We report the experimental results in Table 1.

5. Conclusion

In this paper, we first reported a formula relating the Bayes error B_e (including the probability of error P_e) to the total variation metric TV defined on scaled distributions (see Theorem 1). Second, we elucidated the Chernoff upper bound mechanism based on generalized weighted means: Chernoff [4] used the fact that $\min(a,b) \leqslant a^\alpha b^{1-\alpha}$ for $a, b > 0$ and $\alpha \in [0,1]$ to derive an upper bound that turned out to be well-suited to the structure of exponential families [11]. We interpreted the right-hand side of this inequality as a weighted geometric mean, and considered extending the upper bound construction using generalized weighted mean $M(a,b;\alpha)$. A mean $M(a,b;\alpha)$ is indeed guaranteed to fall within its extrema by definition, thus yielding the bounds: $\min(a,b) \leqslant M(a,b;\alpha) \leqslant \max(a,b)$. We considered the family of quasi-arithmetic means [1,12,14] and showed how to derive new upper bounds by *coupling* the structure of the generalized mean with the structure of the probability distribution family at hand. We illustrated our method by considering three examples: The univariate Cauchy distributions, and the multivariate Pearson type VII and t-distributions (that includes the multivariate normal distributions in the limit case). For those families, we designed new affinity coefficients upper bounding B_e. The best value α^* of α yielding the tightest coefficient can be found by optimization on the statistical manifold [15]. We carried out numerical experiments that show that those novel upper bounds are helpful because not too distant to Bayes error (although not very tight), specially because they can be calculated in constant time using closed-form formula. Otherwise, for more precise approximations, the Bayes error can be estimated using computationally-intensive stochastic integrations.

Last but not least, this paper revealed novel interactions between the Bayes error and statistical divergences: We show how to design $P_e(p_1, p_2) \leqslant \rho_\alpha^f(p_1, p_2)$ upper bounds where ρ_α^f is an affinity coefficient derived from a weighted quasi-arithmetic mean M_f (skewed for Chernoff type and symmetric for a Bhattacharrya type). Since we can transform any affinity coefficient ρ_α^f into a corresponding divergence by defining $D_\alpha^f = -\log \rho_\alpha^f$, we deduce that those novel statistical divergences D_α^f can be used to upper bound the probability of error: $P_e(p_1, p_2) \leqslant e^{-D_\alpha^f(p_1, p_2)}$.

Appendix A. Bayes error for class-conditional Gaussians

We summarize the formula for the Bayes error B_e and the probability of error P_e when dealing with univariate and multivariate normal class-conditional distributions:

$$B_e = \frac{a_1 + a_2}{2} - \text{TV}(a_1 p_1, a_2 p_2), \tag{A.1}$$

$$P_e = \frac{1}{2} - \text{TV}(w_1 p_1, w_2 p_2), \tag{A.2}$$

where $a_1 = w_1(c_{11} + c_{21})$ and $a_2 = w_2(c_{12} + c_{22})$ are the positive weights derived from the cost design matrix and the *a priori* class weights.

First, consider the family of univariate normal distributions $\mathcal{F} = \{N(\mu, \sigma) \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$. Consider two distinct normal densities p_1 and p_2. The two weighted densities $a_1 p_1(x) = a_2 p_2(x)$ intersect at exactly two positions when $\sigma_1 \neq \sigma_2$ or in exactly one position, otherwise:

$$\frac{a_1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - \mu_1}{\sigma_1}\right)^2} = \frac{a_2}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - \mu_2}{\sigma_2}\right)^2}. \tag{A.3}$$

Finding the roots amounts to solve the quadratic equation:

$$\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2 - 2\log\frac{a_1 \sigma_2}{\sigma_1 a_2} = 0. \tag{A.4}$$

When $\sigma_1 = \sigma_2 = \sigma$ (with $\mu_1 \neq \mu_2$), we have one root:

$$x_1 = \frac{\mu_1^2 - \mu_2^2 - 2\log\frac{a_1 \sigma_2}{\sigma_1 a_2}}{2(\mu_1 - \mu_2)}. \tag{A.5}$$

We find:

$$\text{TV}(a_1 p_1, a_2 p_2) = \frac{1}{2} |a_2 \Phi(x_1; \mu_2, \sigma_2) - a_1 \Phi(x_1; \mu_1, \sigma_1)|, \tag{A.6}$$

where $\Phi(x; \mu, \sigma) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right)$ is the cumulative distribution, and $\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$ denotes the error function. That is,

$$B_e = \frac{a_1 + a_2}{2} - \frac{1}{2}\left|a_2 \text{erf}\left(\frac{x_1 - \mu_2}{\sigma\sqrt{2}}\right) - a_1 \text{erf}\left(\frac{x_1 - \mu_1}{\sigma\sqrt{2}}\right)\right|, \tag{A.7}$$

$$x_1 = \frac{\mu_1^2 - \mu_2^2 - 2\log\frac{a_1 \sigma_2}{\sigma_1 a_2}}{2(\mu_1 - \mu_2)}. \tag{A.8}$$

When $\sigma_1 \neq \sigma_2$, the quadratic equation expands as $ax^2 + bx + c = 0$, and we have two distinct roots x_1 and x_2:

$$a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}, \tag{A.9}$$

$$b = 2\left(\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}\right) \tag{A.10}$$

$$c = \left(\frac{\mu_1}{\sigma_1}\right)^2 - \left(\frac{\mu_2}{\sigma_2}\right)^2 - 2\log\frac{a_1 \sigma_2}{a_2 \sigma_1} \tag{A.11}$$

$$x_1 = \frac{-b - \sqrt{\Delta}}{2a}, \quad x_2 = \frac{-b + \sqrt{\Delta}}{2a}, \tag{A.12}$$

with $\Delta = b^2 - 4ac \geqslant 0$ and the total variation writes as follows:

$$TV(a_1 p_1, a_2 p_2) = \frac{1}{2}\left(\left|\text{erf}\left(\frac{x_1 - \mu_1}{\sigma_1 \sqrt{2}}\right) - \text{erf}\left(\frac{x_1 - \mu_2}{\sigma_2 \sqrt{2}}\right)\right| \right.$$
$$\left. + \left|\text{erf}\left(\frac{x_2 - \mu_1}{\sigma_1 \sqrt{2}}\right) - \text{erf}\left(\frac{x_2 - \mu_2}{\sigma_2 \sqrt{2}}\right)\right|\right) \quad (A.13)$$

Those formula generalize the probability of error reported in [3, p. 1375] to the most general case.

Second, consider the family of multivariate normals distributions $\{N(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \succ 0\}$. For densities p_1 and p_2 having the same covariance matrix Σ, the probability of error is reported in [17], even for degenerate covariance matrices Σ by taking the pseudo-inverse matrix Σ^+:

$$P_e = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{1}{2\sqrt{2}}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2 - \mu_1)\|\right). \quad (A.14)$$

When covariance matrices are distinct ($\Sigma_1 \neq \Sigma_2$) but *linear classifiers* are considered, we also get a closed-form formula [18] for the probability of error. Otherwise, for the general case of *quadratic classifiers* of multivariate normals with distinct covariance matrices, no analytical formula is known. The best way to compute the probability of error is then by performing *1D integration* of the *conditional density* of the *discriminant function* [7]. Note that since two matrices can always be simultaneously diagonalized [7], it is enough to consider the case of two Gaussians with the first covariance being set to the identity matrix I and the second covariance matrix set to a diagonal matrix Λ.

References

[1] János D. Aczél, On mean values, Bull. Am. Math. Soc. 54 (4) (1948) 392–400.
[2] Jorge M. Arevalillo, Hilario Navarro, A study of the effect of kurtosis on discriminant analysis under elliptical populations, J. Multivariate Anal. 107 (2012) 53–63.
[3] Chin-Chun Chang, Tzung-Ying Lin, Linear feature extraction by integrating pairwise and global discriminatory information via sequential forward floating selection and kernel QR factorization with column pivoting, Pattern Recognit. 41 (4) (2008) 1373–1383.
[4] Herman Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Stat. 23 (1952) 493–507.
[5] Thomas Cover, Joy A. Thomas, Elements of Information Theory, Wiley-Interscience, 1991.
[6] F. Escolano, P. Suau, B. Bonev, Information Theory in Computer Vision and Pattern Recognition, Springer, 2009.
[7] Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press Professional, Inc., 1990 (first ed., 1972).
[8] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn, mvtnorm: Multivariate Normal and t Distributions, 2013. <http://CRAN.R-project.org/package=mvtnorm>.
[9] Alan Genz, Frank Bretz, Computation of Multivariate Normal and t Probabilities, Lecture Notes in Statistics, 2009.
[10] Martin E. Hellman, Josef Raviv, Probability of error, equivocation and the Chernoff bound, IEEE Trans. Inf. Theory 16 (1970) 368–372.
[11] Thomas Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Trans. Commun. 15 (1) (1967) 52–60.
[12] Andrey Nikolaevich Kolmogorov, Sur la notion de la moyenne, Acad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. 12 (1930) 388–391.
[13] Christophe Ley, Anouk Neven, The normalizing constant in multivariate t-distributions: dimension one versus higher dimensions, Technical report, 2012. arXiv:1211.1174.
[14] Mitio Nagumo, Über eine Klasse der Mittelwerte, Jpn. J. Math. 7 (1930) 71–79 (see Collected papers, Springer 1993).
[15] Frank Nielsen, An information-geometric characterization of Chernoff information, IEEE Signal Process. Lett. 20 (3) (2013) 269–272.
[16] Frank Nielsen, Sylvain Boltz, The Burbea-Rao and Bhattacharyya centroids, IEEE Trans. Inf. Theory 57 (8) (2011) 5455–5466.
[17] Mohammad Hossein Rohban, Prakash Ishwar, Birant Orten, William C. Karl, Venkatesh Saligrama, An impossibility result for high dimensional supervised learning, 2013. arXiv:1301.6915.
[18] Luis Rueda, A one-dimensional analysis for the probability of error of linear classifiers for normally distributed classes, Pattern Recognit. 38 (8) (2005) 1197–1207.
[19] Jianyong Sun, Ata Kabán, Jonathan M. Garibaldi, Robust mixture clustering using Pearson type VII distribution, Pattern Recognit. Lett. 31 (16) (2010) 2447–2454.
[20] Jianxin Wu, James M. Rehg. Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: International Conference on Computer Vision (ICCV), 2009, pp. 630–637.
[21] Alan L. Yuille, James M. Coughlan, Fundamental limits of Bayesian inference: order parameters and phase transitions for road tracking, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2) (2000) 160–173.